A HIERARCHICAL SEGMENTATION ALGORITHM FOR FACE ANALYSIS. APPLICATION TO LIPREADING

M. Liévin and F. Luthon

Signal and Image Laboratory, Grenoble National Polytechnic Institute,
LIS, INPG, 46 av. Félix- Viallet, 38031 Grenoble Cedex, France
email : lievin@lis-viallet.inpg.fr, Franck.Luthon@inpg.fr fax : +33 (0)4 76 57 47 90

ABSTRACT

A hierarchical algorithm for face analysis is presented in this paper. A color video sequence of speaker's face is acquired, under natural lighting conditions and without any particular make-up. The application aims at providing geometrical features of the face for scalable video transmission when no specific model of the speaker face is assumed. First, a logarithmic hue transform is performed from RGB to HI (hue, intensity) color space. Next, a Markov random field modeling regularizes motion and hue information within a spatiotemporal neighborhood. The hierarchical segmentation labels the different areas of the face. Results are shown on the lower part of the face and compared with standard color segmentation algorithm (fuzzy c-means). A speaker's lip shape with inner and outer borders is extracted from the final labeling and used to initialize an active contours stage.

1. INTRODUCTION

The human speech understanding has a bimodal nature (e.g. Benoît in [1]). Actually, visual information can provide a precious help to listener under degraded acoustical conditions or can interfere with the perceived sound (McGurk effect). Audio-visual interface enhances the human-machine interaction in many applications: automatic speech recognition (ASR), synthetic talking faces (videoconference, interactive agent, cartoon animation), communication for disabled people, user verification and recognition (audiovisual biometric features), MPEG4 transmission.

Lipreading is therefore a proper way to improve human-computer interface by using visual features. A typical HCI (Human Computer Interface) scheme is shown in Fig 1. From a video sequence of the speaker, a lipreading algorithm aims at providing a multimedia stream coding a synthetic talking face. The processing is divided into three parts. The first step aims at extracting geometrical face features. Then, an audio processing stage and a morphologic 3-D model compute audio-visual parameters. Finally, the third stage synthesizes a 3-D talking face.

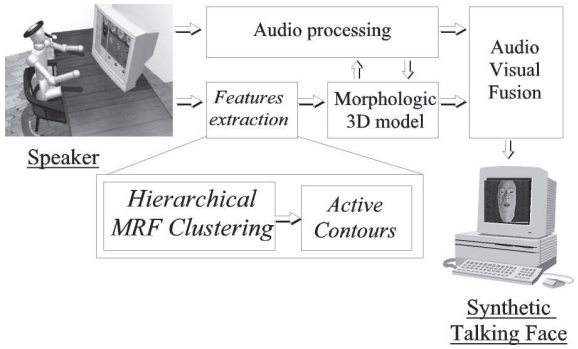

Figure 1: Context of face features extraction.

2. OVERVIEW OF FACE FEATURES EXTRACTION

Because of the unreliable viewing conditions and the various applications, several approaches have been proposed in the literature. The first category uses only the luminance of the image (e.g. Luettin in [8]). In this case, lipreading applications are sensible to the lighting conditions and the region of mouth analysis must be restricted. The second category computes the hue to work in a suitable color space. Indeed, these approaches intend to gain independence from various viewing conditions (e.g. Coianiz in [8]). It appears that color processing is efficient enough to provide robust information for further processing like dynamic contours (e.g. Dalton in [8]).

As said before, lipreading applications are often lacking of well-defined viewing conditions. Moreover, face analysis is easily split up into numerous stages instead of being included in a global scheme. These two reasons often drive the process to require a training stage [3]. Here, we provide a hierarchical algorithm robust to lighting conditions when no specific model of the speaker face is assumed. For that purpose, a specific logarithmic hue model is defined. Next, a statistical algorithm segments the lower part of the face using motion information within a temporal context. Finally, the speaker's lip shape is extracted from the final labeling and used to initialize an active contours stage.

3. LOGARITHMIC COLOR TRANSFORM

To take account of the specific color face distribution, color based approaches often use color angles methods for illuminant invariant recognition. Indeed, color shifts can be well categorized with angles if camera sensors are sufficiently narrowband. The HSI color space is commonly used in face and lip processing.

$$\begin{aligned} I &= \frac{R+G+B}{3} \\ S &= 1 - \frac{\min(R,G,B)}{I} \\ H &= \frac{\pi}{2} - \arctan\left(\frac{2R-G-B}{\sqrt{3}(G-B)}\right) + k \end{aligned} \quad (1)$$

where : $k = 0$ if $G > B$ and $k = \pi$ otherwise.

Unfortunately, a mono-CCD camera gives poor results with angular transforms (noisy conditions). Moreover, G and B channels are correlated in the red region (where R is predominant). From the RGB color space, we use only two components R and G under the assumption that red prevails in face areas and especially in lip areas. The ratio between G and R appears to be more and more used in face and skin processing but rarely justified from a mathematical point of view. We provide here a simplified and justified transform.

To obtain a hue observation robust to lighting conditions, we compute the hue in a mathematical framework based on a logarithmic image processing model [4]. The logarithmic difference becomes a ratio between G and R components when developed at first order. Finally, a simplified logarithmic hue H is defined (Eq. 2).

$$H = \begin{cases} 256 \times \frac{G}{R} & \text{if } G < R, \\ 255 & \text{if } G \geq R. \end{cases} \quad (2)$$

We compute the intensity I as the mean value of the R, G and B components. The figure 2 shows intensity and corresponding hue transforms of five typical images of a speaker sequence : the HSI transform only provides noise where the logarithmic hue reveals the lip shape and face parts.

4. THE SEGMENTATION ALGORITHM

4.1. Hierarchical framework

We consider a hierarchical algorithm in 4 stages, segmenting N regions. At the step $n \in \{1 \dots N\}$, the face segmentation can be summarized as follows (Fig. 3) :

1 : Build a histogram from the hue distribution computed over the non segmented pixels.
2 : Estimate hue cluster vector P_n from zero-crossings of the current histogram derivatives.

Figure 2: *From Top to bottom*: 5 typical images of luminance sequence; the corresponding hue sequence HSI (Eq. 1); the corresponding logarithmic transform (Eq. 2).

3 : Compute hue and motion observations.
4 : Cluster all pixels assigned to the estimated vectors $P_i, i \in \{1 \cdots n\}$ and label them.
If all pixels of the image are not labelled, go to step $n+1$.

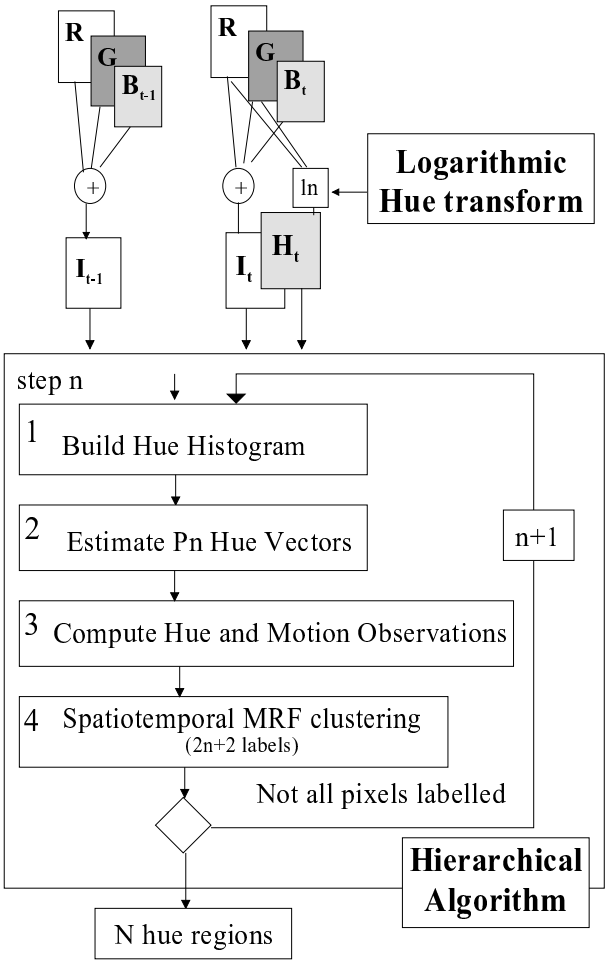

Figure 3: The hierarchical segmentation framework

4.2. Hue parameters estimation

In most of face analysis algorithms, face parameters are determined manually beforehand or by a learning stage. Here, N parameter vectors $P_{n,n\in[1\cdots N]}$ corresponding to N hue regions are automatically estimated. A hue cluster P_n is defined as follow : $P_n = \{H_n, \Delta_{H_n}\}$, where H_n is the mean value of the cluster and Δ_{H_n} is its standard deviation. The following method gives fast results in finding the main mode in a hue distribution P_n :

- Build a smooth histogram with a Gaussian kernel from the hue distribution.
- Compute the first derivative and count the zero-crossings.
- Find the greatest mode H_n corresponding to a zero-crossing.
- Compute the second derivate to estimate Δ_{H_n}.

4.3. Hue and motion observations

To detect face regions, motion information is combined with mean hue. From the HI color space, two kinds of observations are derived, taking values in the same range $[0\cdots 255]$ as the image quantification (8-bit). First, a hue observation $h_n(s)$, corresponding to the n^{th} region, is computed by filtering the hue value $H(s)$ at pixel s with a parabola. This parabola is centered on the mean hue value H_n with a standard deviation of the hue value Δ_{H_n} (Eq. 3). The notation $1_{condition}$ denotes a binary function which takes the value 1 if the condition is true, 0 otherwise.

$$h_n(s) = \left[256 - \left(\frac{H(s) - H_n}{\Delta_{H_n}}\right)^2\right] \times 1_{\frac{|H(s)-H_n|}{\Delta_{H_n}} \leq \sqrt{256}} \quad (3)$$

Second, a temporal observation $fd(s)$ is defined as the unsigned difference between the luminance of two consecutive images (Eq. 4). $I(s)$ represents the intensity (or luminance) at pixel s.

$$fd(s) = |I_t(s) - I_{t-1}(s)| \quad (4)$$

4.4. Labels and MRF framework

The segmentation algorithm requires a relevant hue labeling set corresponding to observations (data set). These labels correspond to a Boolean configuration of motion, N red hue regions and the non-labelled pixels. The label field combining $2N + 2$ labels is then defined at time t. The initial label field L_t^0 is defined by $N + 1$ observations initially thresholded (thresholds θ_h and θ_{fd}).

The algorithm requires an appropriate threshold θ_{fd} to suppress the camera noise without cutting significant temporal changes. We compute here the entropy E_{fd} over an image, $E_{fd} = -\sum_{i\in[0\cdots 255]} p_i log_2(p_i)$ where p_i represents the probability of level i in the observation fd over the image. The threshold motion field is then defined by $fd > \theta_{fd}$ with $\theta_{fd} = 2^{E_{fd}}$. The second threshold θ_h corresponds to the hue observation. We compute $|H(s) - H_i|/\Delta_{H_i}$ within a trust margin of 50% ($\sqrt{256}/2$). The threshold hue field is then defined by $h_i > \theta_h$ with $\theta_h = 192$. The threshold fields are of course non homogeneous and noisy. Therefore, a statistical relaxation is needed to segment more accurately the lips.

We apply in each step of the hierarchical algorithm the MRF (Markov Random Field) framework defined in the previous work [6]. This field maximizes the posterior probability of the labeling in a specific spatiotemporal neighborhood [2]. In the spatiotemporal neighborhood structure, elementary potentials are defined to constrain the model to spatial and temporal homogeneity.

5. EXPERIMENTAL RESULTS

5.1. Face segmentation

The first application is to provide accurate areas detection for skin texture extraction and relevant regions location (face, shirt, background). In order to evaluate the efficiency of the segmentation, we compare our results with the color fuzzy C-means algorithm (FCM) [7]. Since the later is not dedicated to specific viewing conditions, the face can easily be segmented but few regions can be extracted inside. Moreover, this algorithm requires specific parameters for each image: the lips disappear on the second image of the figure 4. On the other hand, our hierarchical algorithm extracts lip areas accurately and automatically. The labels correspond to the shirt, the face skin, the lips and the inner mouth. Thinner regions correspond to moving areas. The parameter estimation algorithm is robust enough to adapt to different speakers without manual adjustments (Fig. 5).

Figure 4: Results of final labeling with constant parameters. From Top to Bottom : the luminance sequence; the corresponding fuzzy-c means labeling; the corresponding hierarchical labeling.

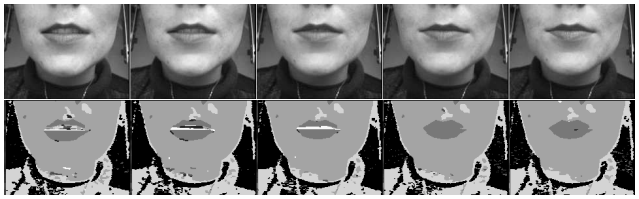

Figure 5: Results of hierarchical segmentation in the case of a speaker with a red make-up and a soft additive light.

5.2. Application for Lipreading

The segmentation algorithm detailed in this paper can be used for lipreading. A preprocessing stage consists in locating and segmenting the lips in adverse viewing conditions (no illumination and no specific make-up). The second stage consists in an active contours processing to detect precisely the borders and the corners of the lips.

Two steps of the hierarchical approach are required to extract the areas of the lips: the first step segments the face skin; the second step corresponds to the lip hue areas extraction. The hue estimation process is then defined as:

- **Step** 1 : Cluster the main hue mode P_1 which corresponds to the face skin of the speaker.
- **Step** 2 : Cluster the second mode P_2 with the constraint $H_2 < H_1$ (the lip hue mode has its mean lower than the face skin mode : the lips are more red than the face).
- Keep the labels corresponding the second step in the hierarchical segmentation (the lip hue mode).

Finally, the hierarchical algorithm extracts the speaker's lip shape. The labeling is robust enough to be used for initializing a further processing, namely an active contours extraction [5] (Fig. 6).

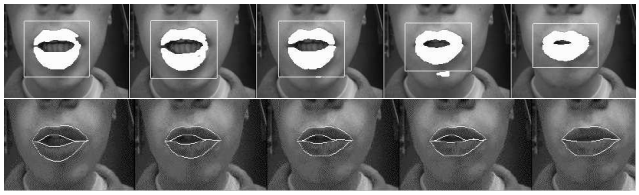

Figure 6: A sequence of lip hue segmented with ROI superimposed on the corresponding luminance and the active contours results.

6. CONCLUSION AND FORTHCOMING DEVELOPMENTS

A hierarchical color segmentation algorithm has been successfully applied to several sequences when no specific model of the speaker and with variable viewing conditions. A logarithmic color transform followed by a spatiotemporal hierarchical segmentation dealing with hue and motion information segment each region of the face. Then, various postprocessing can be applied depending on the applications: texture extraction, active contours...

The proposed algorithm requires less than 2 seconds per image 256×256 and per stage on a standard 200 MHz workstation (10 sec. per image when labeling 5 regions). Therefore, hardware implementation to reach video rate is currently under study. An application for lip segmentation has been successfully implemented for geometrical lip features extraction whereas classic segmentation algorithms, namely fuzzy-c means, like manual fine-tuning.

We are still not able to detect precisely the borders of the lips in areas where tongue or gum appear. These two features do not differ from the lip area as regards hue or motion, so we need to introduce another observation (gradients, texture).

References

[1] C. Benoît, M.T. Lallouache, and T. Mohamadi. A set of French visemes for visual speech synthesis. In *Talking Machines: Theories, Models and Designs*, pages 485–504. Elseviers Science Publishers, 1992.

[2] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Int.*, 6(6):721–741, November 1984.

[3] H.P. Graf, T. Chen, E. Petajan, and E. Cosatto. Locating faces and facial parts. In *International Workshop on Automatic Face and Gesture Recognition*, pages 41–44, Zurich, 1995.

[4] M. Jourlin and J-C. Pinoli. Image dynamic range enhancement and stabilization in the context of the logarithmic image processing model. *Signal Processing*, 41(2):225–237, January 1995.

[5] M. Liévin, P. Delmas, P.Y. Coulon, F. Luthon, and V. Fristot. Automatic lip tracking : Bayesian segmentation and active contours in a cooperative scheme. In *Proc. of IEEE Int. Conf. on Multimedia, Computing and Systems*, Florence, Italie, June 1999.

[6] M. Liévin and F. Luthon. Unsupervised lip segmentation under natural conditions. In *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 6, pages 3065–3068, Phoenix, Arizona, March 1999.

[7] Y.W. Lim and S.U. Lee. On the color image segmentation algorithm based on the thresholding and the fuzzy c-means techniques. *Pattern Recognition*, 23(9):935–952, 1990.

[8] D. Stork and M. Hennecke. *Speechreading by Humans and Machines*, volume 150. Springer-Verlag, Berlin, 1996.

Acknowledgments

We thank P. Delmas and P.Y. Coulon for fruitfully collaboration and results in active contours applications [5] (Fig. 6).